

REGULARIZATION TECHNIQUES FOR PSF-MATCHING KERNELS. I. CHOICE OF KERNEL BASIS

A.C. BECKER¹, D. HOMRIGHAUSEN², A.J. CONNOLLY¹, C.R. GENOVESE², R. OWEN¹, S.J. BICKERTON³, R.H. LUPTON³*Draft version February 15, 2012*

ABSTRACT

We review current methods for building PSF-matching kernels for the purposes of image subtraction or coaddition. Such methods use a linear decomposition of the kernel on a series of basis functions. The correct choice of these basis functions is fundamental to the efficiency and effectiveness of the matching – the chosen bases should represent the underlying signal using a reasonably small number of shapes, and/or have a minimum number of user-adjustable tuning parameters. We examine methods whose bases comprise multiple Gauss-Hermite polynomials, as well as a form free basis composed of delta-functions. Kernels derived from delta-functions are unsurprisingly shown to be more expressive; they are able to take more general shapes and perform better in situations where sum-of-Gaussian methods are known to fail. However, due to its many degrees of freedom (the maximum number allowed by the kernel size) this basis tends to overfit the problem, and yields noisy kernels having large variance. We introduce a new technique to regularize these delta-function kernel solutions, which bridges the gap between the generality of delta-function kernels, and the compactness of sum-of-Gaussian kernels. Through this regularization we are able to create general kernel solutions that represent the intrinsic shape of the PSF-matching kernel with only one degree of freedom, the strength of the regularization λ . The role of λ is effectively to exchange variance in the resulting difference image with variance in the kernel itself. We examine considerations in choosing the value of λ , including statistical risk estimators and the ability of the solution to predict solutions for adjacent areas. Both of these suggest moderate strengths of λ between 0.1 and 1.0, although this optimization is likely dataset dependent. This model allows for flexible representations of the convolution kernel that have significant predictive ability, and will prove useful in implementing robust image subtraction pipelines that must address hundreds to thousands of images per night.

Subject headings: methods: data analysis, techniques: image processing, techniques: photometric

1. INTRODUCTION

Studies of variability in astronomy typically use image subtraction techniques in order to characterize the magnitude and type of the variability. This practice involves subtracting a prior-epoch (generally high signal-to-noise) template image from a recent science image; any flux remaining in their difference may be attributed to phenomena that have varied in the interim. This technique is sensitive to both photometric and astrometric variability, and can uncover variability of both point-sources (such as stars or supernovae; e.g. Udalski et al. 2008; Sako et al. 2008) and extended-sources (such as comets or light echoes; e.g. Newman & Rest 2006). Successful application of this technique shows that it is sensitive to variability at the Poisson noise limit in a variety of astrophysical conditions (Alard & Lupton 1998; Alard 2000; Bramich 2008; Kerins et al. 2010), and in this regard may be considered optimal.

There are several reasons for preferring such an approach over catalog-based searches. First, many types of variability are found in confused regions of the sky, and it may be difficult to deblend the time-variable signal from the non-temporally-variable surrounding area. This is particularly true for supernovae and active galac-

tic nuclei, which are typically blended with light from their host galaxies. However, such confusion is not limited to stationary objects. Moving solar system bodies may serendipitously yield false brightness enhancements in the measurement of a background object if the impact parameter is small compared to the image's point spread function (PSF). For this reason, removal of non-variable objects is preferred before attempting to characterize variable sources in images.

Image subtraction is also an efficient technique as the vast majority of pixels in an image do not contain signatures of astrophysical variability. Any pixel-level analysis of a difference image will, therefore, be restricted to those sources that are temporally variable (as opposed to analyzing all sources within an image). While many variants of this technique have been published (Tomaney & Crotts 1996; Alard & Lupton 1998; Bramich 2008; Albrow et al. 2009), and many versions implemented in automated variability-detection pipelines (Bond et al. 2001; Rest et al. 2005; Darnley et al. 2007; Miller et al. 2008; Udalski et al. 2008), there does remain room for improvement in the robustness of the image subtraction, and in the reduction of subtraction artifacts. We refer the reader to Wozniak (2008) for an in-depth summary on the practical application of these image subtraction techniques.

¹ Astronomy Department, University of Washington, Seattle, WA 98195

² Carnegie Mellon University, Department of Statistics. Pittsburgh, PA 15232

³ Department of Astrophysical Sciences, Princeton University, Princeton, NJ 08544

In image subtraction we assume that we have two images of the same portion of the sky, taken at different epochs, but in the same filter. We will call the image that contains the variability of interest the “science” image, and the template image to be subtracted the “reference” image. The images will, in general, be astrometrically misaligned, but this can be resolved by using sinc-based image registration methods that preserve the noise properties of the original image. After astrometric alignment, a given astrophysical object will be represented in the reference image as a sub-array of pixels $R(x, y)$ and in the science image as $S(x, y)$, with the same span in x and y . Each image will, however, have a different point spread function (PSF), which is the spatial response of a point source due to the atmosphere, telescope optics, and instrumental signatures. PSF-matching of the images is required before we can subtract one image from the other, and is the essence of the image subtraction technique.

2. PSF-MATCHING

We typically assume that the reference image is a high signal-to-noise (S/N) representation of the field, for example an image coadd made through a mosaicing process, or a single image taken on a night with particularly good seeing. A standard assumption (e.g. Alard & Lupton 1998; Alard 2000) is that $S(x, y)$ can be modeled as a convolution of $R(x, y)$ by a single PSF-matching kernel $K(u, v)$, with an additional noise component $\epsilon(x, y)$:

$$S(x, y) = (K \otimes R)(x, y) + \epsilon(x, y). \quad (1)$$

Our goal in this paper is to develop an effective method for determining $K(u, v)$.

2.1. Linear Modeling of $K(u, v)$

As inputs to the PSF-matching technique, we assume images are astrometrically registered, and background subtracted (while this latter constraint is not a necessity, it does enable us to restrict our analysis here to the respective shapes of the PSFs). To proceed, we make the assumption that $K(u, v)$ may be modeled as a linear combination of basis functions $K_i(u, v)$, such that $K(u, v) = \sum_i a_i K_i(u, v)$ (Alard & Lupton 1998). The basis components do not have to be orthonormal, nor does the basis need to be complete (indeed, it may be overcomplete). However, it is desirable to choose a shape set that compactly describes $K(u, v)$, such that the number of required terms is small.

By formulating the kernel decomposition as a linear expansion, we may recast Equation 1 as the vectorized equation

$$S = Ca + \epsilon \quad (2)$$

where C is the matrix of functions $C_i \equiv (K_i \otimes R)$ evaluated at each pixel. For any given kernel basis set, the goal is to find the coefficients a_i associated with each K_i .

We proceed using standard linear least squares analysis. We assume that the noise is uncorrelated and known; ϵ is therefore the product of a diagonal matrix $\Sigma^{1/2}$, which represents the square root of the known per-pixel variance, and a zero-mean unit-variance random variable Z . By reweighting by the inverse square root of Σ

(which must exist as covariance matrices are positive definite and hence invertible) we obtain the modified equation

$$\Sigma^{-1/2}S = \Sigma^{-1/2}Ca + Z \quad (3)$$

which is just another linear model, now with the error term Z having the identity matrix for the covariance. This reduces to the weighted linear least squares equation

$$\tilde{S} = \tilde{C}a + Z, \quad (4)$$

with

$$\begin{aligned} \tilde{S} &\equiv \Sigma^{-1/2}S, \\ \tilde{C} &\equiv \Sigma^{-1/2}C. \end{aligned} \quad (5)$$

The normal equations for estimating a are:

$$\begin{aligned} \tilde{C}^\top \tilde{S} &= \tilde{C}^\top \tilde{C}a \\ C^\top \Sigma^{-1}S &= C^\top \Sigma^{-1}Ca. \end{aligned} \quad (6)$$

This may be cast in the familiar form of $b = Ma$ with

$$\begin{aligned} b &= C^\top \Sigma^{-1}S \\ M &= C^\top \Sigma^{-1}C. \end{aligned} \quad (7)$$

In discrete pixel coordinates, this corresponds to

$$\begin{aligned} b_i &= \sum_{x,y} \frac{C_i(x, y)S(x, y)}{\sigma^2(x, y)} \\ M_{ij} &= \sum_{x,y} \frac{C_i(x, y)C_j(x, y)}{\sigma^2(x, y)} \end{aligned} \quad (8)$$

where $\sigma^2(x, y)$ represents the known variance per pixel. The creation of the matrices M_{ij} and b_i therefore requires a convolution of the reference image with each basis kernel.

The least-squares estimate for a is $\hat{a} = (\tilde{C}^\top \tilde{C})^{-1} \tilde{C}^\top \tilde{S}$. A difference image is then constructed as $D(x, y) = S(x, y) - C\hat{a}(x, y)$. Because the estimate of \hat{a} is explicitly dependent on both $S(x, y)$ and $R(x, y)$, the residuals in the difference image may *not* necessarily follow a normal $\mathcal{N}(0, 1)$ distribution⁴, with $\sigma^2 \neq 1$ due to this covariance. The residuals should however have a flat power spectral density.

2.2. Invertability of $C^\top \Sigma^{-1}C$

When a large set of basis functions is used, the matrix $M = C^\top \Sigma^{-1}C$ may be ill-conditioned or even singular. This can be quantified by the “condition number” of M , which we define as the ratio of the largest to the smallest eigenvalues. When the condition number is large, inversion of M will be numerically unstable or infeasible.

A common approach when trying to invert an ill-conditioned matrix is to compute instead a *pseudo-inverse*, or an approximation to one in which eigenvalues that are numerically small are zeroed out. As M is symmetric, we can decompose it as $M = VDV^\top$ with V an orthogonal matrix and $D = \text{diag}(d_1, \dots, d_n)$ with eigenvalues $d_1 \geq d_2 \geq \dots \geq d_n \geq 0$. We define $D_i =$

⁴ We use the mean and variance, not mean and standard deviation, as the two parameters of Normal distributions.

$\text{diag}(d_1, \dots, d_i, 0, \dots, 0)$ to be a truncation of D where d_1/d_{i+1} becomes too large. Then, we define the pseudo-inverse of D_i as $D_i^\dagger = \text{diag}(1/d_1, \dots, 1/d_i, 0, \dots, 0)$. Note this allows for the definition of a pseudo-inverse of M as $M^\dagger = VD_i^\dagger V^T$. Analogous to D_i , define V_i to be the same as the matrix V in the first i columns, and zero elsewhere. Typically this truncation threshold is defined by the machine precision of the computation (e.g. for double-valued calculations, $1/d_{\min} \sim 5 \times 10^{15}$). However, significantly larger limits for d_{\min} may be used to avoid underconstrained parameters, such as in Section 5.1.1.

3. SUM-OF-GAUSSIAN BASES

The original PSF-matching bases proposed by Alard & Lupton (1998) and Alard (2000) (referred to here as “Alard-Lupton” or AL bases) used a sum of multiple Gaussians, each modified by a 2-dimensional polynomial:

$$K_i(u, v) = e^{-(u^2+v^2)/2\sigma_n^2} u^p v^q, \quad (9)$$

where the index i runs over all permutations of n, p, q . This basis set effectively uses $n = 1 \dots N$ Gaussian components, each with width σ_n , and each modified by a set of Gauss-Hermite polynomials (e.g. Wnsche 2000) expanded out to order O_n ($0 \leq p + q \leq O_n$). The total number of basis functions in the set is $\sum_n (O_n + 1) \times (O_n + 2)/2$.

The number N and width σ_n of the Gaussians, as well as spatial order of the polynomials O_n , are configurable but are not fitted parameters in the linear least-squares minimization. Therefore these are tuning parameters of the model. Typically, *a-priori* information such as the widths of the image PSFs is used to choose these values (e.g. Israel et al. 2007). In a representative implementation (Smith et al. 2002), three Gaussians are used, with the narrowest Gaussian expanded out to order 6, the middle to order 4, and the widest to order 2. This leads to a total of 49 basis functions used in the kernel expansion.

The practical application of this algorithm has been very successful, and it has been used by various time-domain surveys such as MACHO (Alcock et al. 1999), OGLE (Wozniak 2000; Udalski et al. 2008), MOA (Bond et al. 2001), SuperMACHO (Smith et al. 2002; Rest et al. 2005), the Deep Lens Survey (Becker et al. 2004), ESSENCE (Miknaitis et al. 2007), the SDSS-II Supernova Survey (Sako et al. 2008), and most recently analysis of commissioning data from Pan-STARRS (Botticella et al. 2010).

The top row of Figure 1 shows an instance of successful PSF matching using this sum-of-Gaussians basis. The first column represents a high signal-to-noise image of a star $R(x, y)$ generated from an image coaddition process applied to data from the Canada-France-Hawaii Telescope (CFHT). The second column shows this same star, aligned with the template image to sub-pixel accuracy, in a single science image $S(x, y)$. The star is obviously asymmetric, potentially due to optical distortions such as focus or astigmatism, or due to tracking problems during acquisition of the image. The PSF-matching kernel thus will need to take the symmetric $R(x, y)$ and elongate it along a vector oriented approximately 135 deg from hor-

izontal. The first row, third column shows the best-fit PSF-matching kernel using $N = 3$ Gaussians with $\sigma_n = [0.75, 1.5, 3.0]$ pixels, and each modified by Hermite polynomials of order $O_n = [4, 3, 2]$, respectively. The total number of terms in the expansion is 31. The first row, right column shows the resulting difference image $D(x, y)$. The subtraction is obviously very good, with the remaining pixels described by a $\mathcal{N}(0.01, 1.01)$ distribution.

3.1. Limitations of the Model

The intrinsic symmetries of Hermite polynomials (symmetric for even order, anti-symmetric for odd order) means that the Gauss-Hermite bases possess a high degree of symmetry about the central pixel. This makes it difficult to concentrate the kernel power off-center when using an incomplete basis expansion. Such functionality is necessary when the flux needs to be redistributed on the scale of the kernel size, such as when there are astrometric misalignments. While it is possible to compensate for misalignment using kernels derived from this basis, this requires concentrating the kernel strength in the high-order terms. There are practical limitations to the efficacy of this including the scale and orientation of the required shift, and the number of basis terms used.

As a concrete example, the second row in Figure 1 shows the best-fit kernel derived when there is a 3-pixel shift in both the x and y directions. The kernel needs to have power in the first quadrant (upper right) at the scale of 3 pixels. The image of the kernel (third column) shows that while it is obviously able to do so, the matching suffers in the third quadrant, as the difference image shows obvious residuals. These pixels result in an unacceptable $\mathcal{N}(0.01, 1.44)$ distribution; recall we were able to yield $\sigma^2 = 1.01$ for well-registered images (top row).

Another limitation of the model is that there are a variety of tuning parameters. This includes the number of Gaussians in the basis, their widths, and their spatial orders. These parameters are typically chosen using a set of heuristics. If there is a mismatch compared to the true underlying kernel, this process will fail. The third row of Figure 1 shows PSF-matching results when the basis Gaussians are *too big* and are unable to reproduce the small-scale differences in the PSFs. This yields obvious residuals in the difference image, which follow a $\mathcal{N}(0.02, 3.03)$ distribution. The fourth row of Figure 1 shows results when the Gauss-Hermite polynomials are not allowed to vary to high enough order, also yielding unacceptable $\mathcal{N}(0.02, 2.86)$ residuals in the difference image.

Clearly the results of this process are sensitive to the choice of several tuning parameters, which makes this difficult to implement robustly. In a statistical sense, selection of tuning parameters (which includes selecting the number of basis functions used) usually has a much larger effect on performance than does the choice of basis functions. A process that results in a reduction in the number of kernel tuning parameters, while maintaining the quality of the difference images, would greatly improve the effectiveness of this method.

4. DELTA-FUNCTION BASES

The most general technique for modeling $K(u, v)$ is to use a “shape free” basis, which consists of a delta

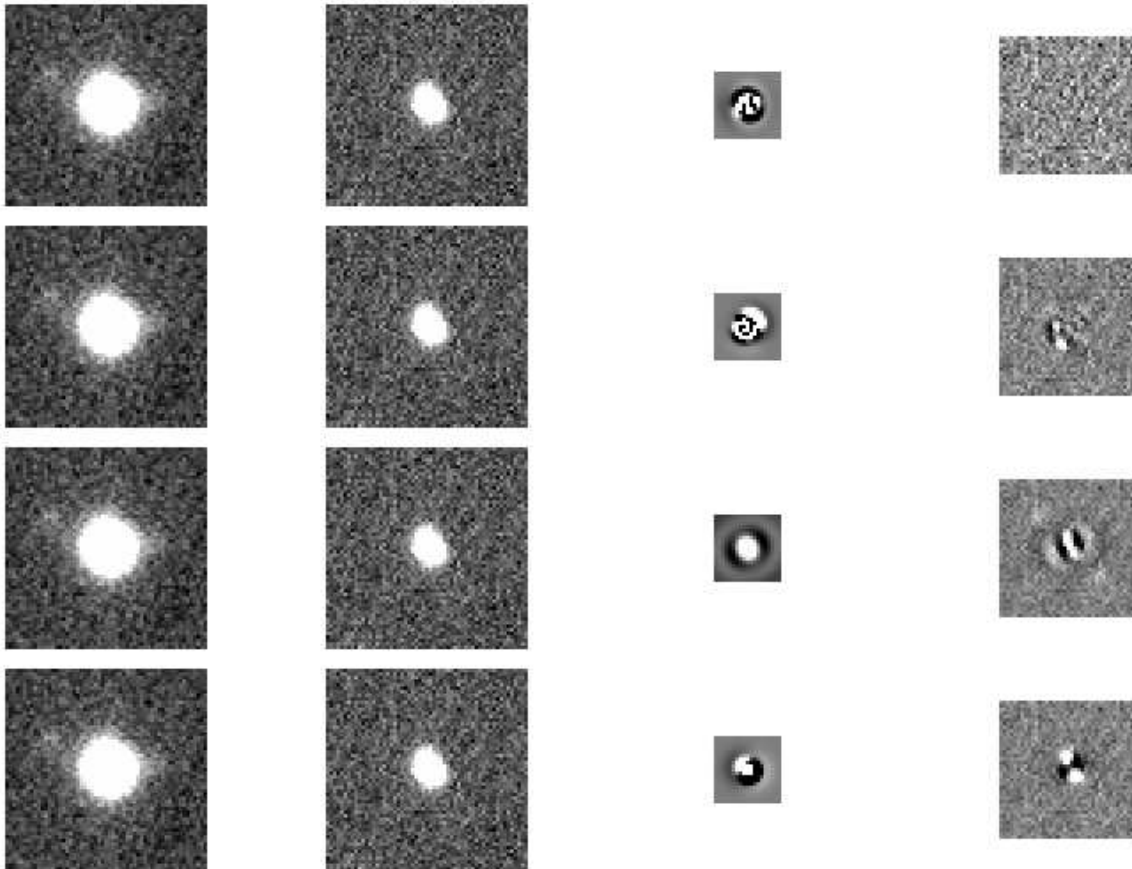


Figure 1. Difference imaging results when using a sum-of-Gaussian basis. The first column shows the reference image to be convolved $R(x, y)$, the second shows the science image $S(x, y)$ the reference is matched to, the third column shows the best-fit 19×19 pixel PSF matching kernel $K(u, v)$, and the fourth column shows the resulting difference image $D(x, y)$. *Row 1:* Results when using a basis set with $\sigma_n = [0.75, 1.5, 3.0]$ pixels, $O_n = [4, 3, 2]$. *Row 2:* Results when the images are mis-registered by 3 pixels in both coordinates, requiring significant off-center power in the kernel. *Row 3:* Results when the basis Gaussians are too large compared to the actual PSF-matching kernel ($\sigma_n = [3.0, 5.0]$ pixels, $O_n = [3, 2]$). *Row 4:* Results when the polynomial expansion is not carried to high enough order ($\sigma_n = [0.75, 1.5, 3.0]$ pixels, $O_n = [1, 1, 1]$).

function at each kernel pixel index: $K_{ij}(u, v) = \delta(u - i)\delta(v - j)$. A kernel of size 19×19 will then have 361 orthonormal, single-pixel bases. In this situation there are *no* tuning parameters, which is an obvious benefit. However, in any choice of basis there is a trade off between flexibility in the forms the fitted function can take, and variability in the resulting fit (the so-called “bias-variance” trade off). The delta-function basis provides complete flexibility, and as such can account for features such as arbitrary off-center power required to compensate for astrometric misregistration (e.g. Bramich 2008). But to avoid gross overfitting, that flexibility needs to be tempered to keep the variance in check.

Figure 2 shows the results of PSF-matching using such a basis, using the same objects as in Figure 1. The top row demonstrates the results for exactly aligned images, while the bottom row demonstrates the results for images misaligned by 3 pixels in both x and y . The difference images are qualitatively similar. However, the best-fit solutions obviously yield large variations within the kernels themselves, and do not match expectations of what the actual kernel should look like. The reason for this can be found in the distribution of pixels residuals in the difference image. Both images follow a $\mathcal{N}(0.01, 0.79)$

distribution. This indicates that the residuals have lower variance than Gaussian statistics would suggest. Indeed, in Figure 2 column 4 the residuals appear smoother than random noise. This is impossible unless we have overestimated the variance in our images, or unless the kernels themselves are removing some fraction of the noise.

The large numbers of basis shapes (361 degrees of freedom vs. 31 for the sum-of-Gaussians) makes it highly likely that we are over-fitting the problem. The kernel thus has the ability to match both the underlying signal *and* the associated noise in the two images. So while this technique is optimal for matching pixels in two images – where those pixels are a combination of signal and noise – it is not necessarily optimal for uncovering the true PSF-matching kernel.

A consequence of this is that the PSF-matching kernel derived for any given object may not be directly applied to neighboring objects, since the solution is significantly driven by the local noise properties. High variance estimators are particularly poor as inputs to interpolation routines, or to a spatial model of the kernel $K(u, v, x, y)$, that find the matching kernel at *all* locations as a function of the fitted kernels at particular locations. Below, we explore how introducing a certain amount of bias into

this estimator can improve its performance.

5. DELTA-FUNCTION BASES WITH REGULARIZATION

The delta-function basis can flexibly fit a kernel of any form, but as we have shown, this flexibility is both its strength and weakness. As is, the method significantly overfits, absorbing substantial noise fluctuations into the fit and thus giving estimated kernels with excessive variance. A solution is to introduce some amount of bias into the fit to reduce the solution variance by a much larger factor (if "bias" sounds pejorative, note that this is just a kind of smoothing). When fitting a smooth function such as $K(u, v, x, y)$, we prefer fitted kernels for which nearby solutions do not vary too greatly. This bias will enable such a fit with vastly reduced mean-squared error.

Among the various approaches to dealing with overfitting, the most common are through linear regularization techniques (e.g. Section 18.5; Press et al. 1992). Using these, we may penalize undesirable features of the fit, usually by adding a penalty term to our optimization criterion. For instance, when fitting a smooth function, we want to penalize fits f that are too rough or irregular. One way to do this is to add to the least squares objective a term penalizing the second derivative, $\lambda \cdot \int |f''(x)|^2 dx$. Here, the scaling factor λ is a tuning parameter that determines the balance between fidelity to the data and the desired smoothness. In the case of kernel matching, we may extend this idea with a two dimensional penalty that approximates $\lambda \cdot \int \int |\nabla f(x, y)|^2 dx dy$.

The one-dimensional second derivative of a function f around pixel x may be approximated using the central finite difference $f''(x) \approx f(x-1) - 2f(x) + f(x+1)$. Since the delta-function bases have unit height and no intrinsic shape, regularizing the coefficients a_i is equivalent to regularizing the shape of the resulting kernel (care must be taken to apply the regularization penalty to only those pixels that are associated spatially). In matrix terms, this one-dimensional regularization may be represented by $R_1 a$, with

$$R_1 = \begin{pmatrix} 1 & -2 & 1 & 0 & \dots & 0 \\ 0 & 1 & -2 & 1 & \dots & 0 \\ \vdots & & & & & \vdots \\ 0 & \dots & 0 & 1 & -2 & 1 \end{pmatrix} \quad (10)$$

which is of dimension $(m-2) \times m$, where m here is the total number of pixels in the kernel⁵. A generalization of this to two dimensions results in a 5-point stencil that sums the local derivative along both axes, $f''(x, y) \approx f(x-1, y) + f(x+1, y) + f(x, y-1) + f(x, y+1) - 4f(x, y)$, with an associated matrix R_2 .

The finite calculation of this penalty is implemented through the matrix equations

$$|R_2 a|^2 = a^\top R_2^\top R_2 a \quad (11)$$

where a represents the amplitude of each delta function, and R_2 encapsulates the coefficients that approximate the local 2nd derivative of the resulting kernel. We define the matrix $H \equiv R_2^\top R_2$, which makes the second derivative penalty $a^\top H a$. This matrix is used to regularize the

⁵ The absolute value of the kernel's border pixels may also be penalized through the addition of a row at both the top and bottom of R_1 .

normal equations (Equation 6) with strength λ

$$\begin{aligned} C^\top \Sigma^{-1} S &= (C^\top \Sigma^{-1} C + \lambda H) a \\ b &= M_\lambda a. \end{aligned} \quad (12)$$

Note the similarity to Equation 4, with the only difference being $M_\lambda = M + \lambda H$. Here λ represents the strength of the regularization penalty, and is the sole tuning parameter in this model.

Figure 3 shows results for the same set of objects displayed in Figure 1 and Figure 2, but using regularization of the delta-function basis set. The top row shows the results for aligned images, and $\lambda = 1$. Note that the kernel looks very much as anticipated, being compact and having a shape aligned approximately 135 deg from horizontal. Residuals in the difference image follow a $\mathcal{N}(0.01, 0.94)$ distribution. The second row shows the results when the images are misaligned by 3 pixels in x and y . The kernel merely appears shifted by the same amount compared to the aligned images, and the difference image follows a quantitatively similar $\mathcal{N}(0.01, 0.96)$ distribution. This effectively demonstrates that this method can reproduce kernels with off-center power. The third row shows the results with $\lambda = 0.01$; the shape of the PSF-matching component of the kernel is just barely discernible above its noise, suggesting the regularization is too weak. The difference image is, however, acceptable ($\mathcal{N}(0.01, 0.81)$). The fourth row shows the results with $\lambda = 100$. The kernel is far smoother than in previous runs. However, this appears to be at the expense of residuals in the difference image, which follow a $\mathcal{N}(0.01, 1.35)$ distribution. This suggests that too much weight has been given the smoothness of the kernel compared to the residuals in the difference image, indicating that the regularization is too strong. The general trend is that with increasing lambda, the variance in the difference image increases. The noise properties of the difference image evolve from being too smooth, to approximately white in spectrum, to having residual features at a similar scale as the kernel.

Overall, this technique appears very effective. We are able to create general, compact kernels that represent the underlying shape of the PSF-matching kernel with only one tuning parameter, the strength of the regularization λ . The role of λ is effectively to exchange variance in the resulting difference image with variance in the kernel itself. By increasing the value of λ , we are able to smooth the kernel while increasing the variance in the difference image. We explore various methods to establish the optimal value of λ below.

5.1. Choice of Tuning Parameter λ

Choosing a good tuning parameter is essential for good performance of a regularization method. If λ is too high, the fit will be too smooth (high bias, low variance); if λ is too low, the fit will be too rough (low bias, high variance). The goal of data-driven methods for choosing tuning parameters is to find the sweet spot in the bias-variance trade off. While choosing a good value for λ is a hard statistical problem, there are a variety of methods that have proven successful in practice. These methods construct a statistical estimate of mean-squared error and choose λ to minimize it. For instance in cross-validation (reviewed in Kohavi 1995), the data set is broken into

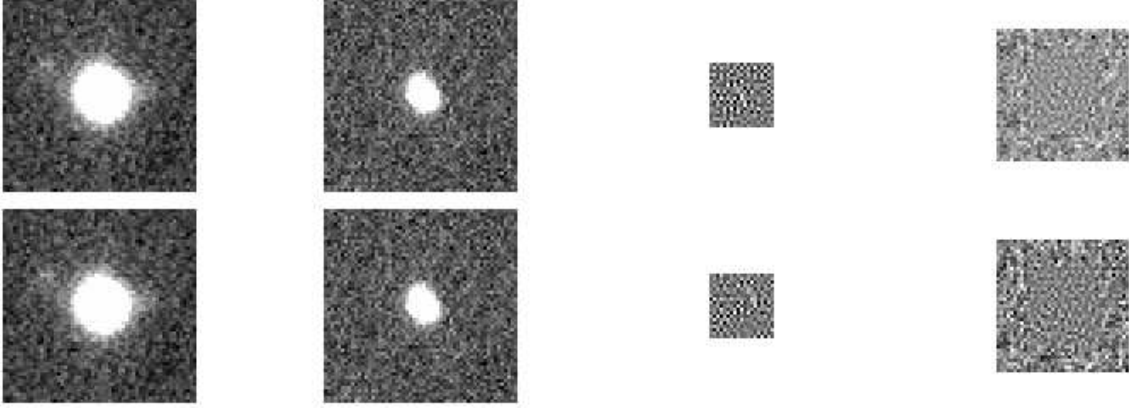


Figure 2. Difference imaging results when using a delta-function basis. Columns are the same as in Figure 1. *Row 1:* Results when using an unregularized delta-function basis. *Row 2:* Results when the images are mis-registered by 3 pixels in both coordinates.

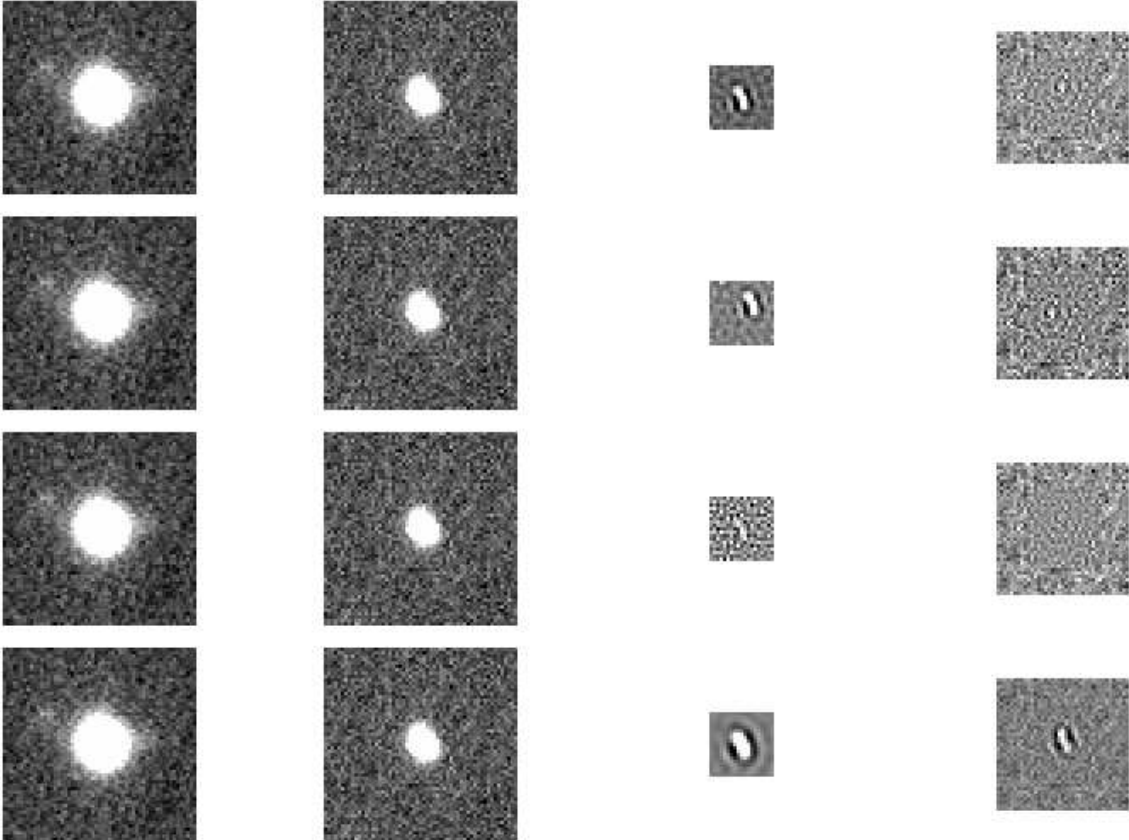


Figure 3. Difference imaging results when using a regularized delta-function basis. Columns are the same as in Figure 1. *Row 1:* Results when using a regularized delta-function basis with $\lambda = 1.0$. *Row 2:* Results when the images are mis-registered by 3 pixels in both coordinates, $\lambda = 1.0$. *Row 3:* Results using “weak” regularization of the kernel, with $\lambda = 0.01$. *Row 4:* Results using “strong” regularization of the kernel, with $\lambda = 100$.

pieces, and each piece is left out in turn during the fit. The (prediction) mean-squared error is derived from the average squared error of the fits in predicting the part of the data that was left out. Another approach, called empirical risk estimation (Stein 1981), uses the data itself to compute an (unbiased) estimate of original fit’s mean-squared error and chooses λ to minimize it. The theoretical justification for these methods is that, when

properly done and with sufficiently large data sets, the chosen λ is close to the value that minimizes the corresponding mean-squared error function.

A second tuning consideration is that frequently a set of fitted kernels will be used to constrain a spatial model $K(u, v, x, y)$ that will be applied to *all* pixels in an image. Therefore we must give a large weight to our ability to interpolate between the ensemble of kernel realizations

used to constrain $K(u, v, x, y)$. One metric for this is to examine the predictive power of a kernel derived from one object, and applied to a neighboring object. At small separations, the quality of each difference image should be similar, indicating that the initial solution was not significantly driven by the local noise properties.

We explore the practical application of these ideas below using several sets of CCD images from the Canada–France–Hawaii Telescope plus Megacam imager, calibrated using the ELIXIR pipeline of Magnier & Cuillandre (2004). The template image used is the median of several images into a single high S/N representation of the field. The variance per pixel is determined from the image pixel values divided by the gain.

5.1.1. Empirical Risk Estimation

We first construct a loss function that represents the sum of squared differences between the true (unknown) kernel coefficients a and \hat{a}_λ , which is our estimate of the kernel coefficients when the tuning parameter is set to the value λ :

$$L(a, \hat{a}_\lambda) = (\hat{a}_\lambda - a)^\top (\hat{a}_\lambda - a). \quad (13)$$

The expectation value of $L(a, \hat{a}_\lambda)$ is the statistical risk we will minimize through our choice of λ ⁶. When M is well-conditioned, we can construct an unbiased estimator of the true risk $\langle L(a, \hat{a}_\lambda) \rangle$ as (Section 2, Stein 1981)

$$R(\lambda) = \|\hat{a}_\lambda\|_2^2 + 2(\text{trace}(M_\lambda) - \hat{a}_\lambda^\top M^{-1}b). \quad (14)$$

We note that this estimator of risk does not require tuning parameters. If we let $\hat{\lambda}$ be the minimizer of R , then we choose $\hat{a}_{\hat{\lambda}}$ as the estimate of a .

For the circumstance that M is ill-conditioned, we present an adjustment to R from Equation 14. Following the notation from Section 2.2, for any i define $\Pi = V_i V_i^\top$. This corresponds to Π being a projection matrix onto the space of the eigenvectors of M that correspond to its i largest eigenvalues. Note that i is now an additional tuning parameter, corresponding to choice of condition number (denoted by symbol Λ) for matrix M (Section 2.2). A biased estimate of the statistical risk is then:

$$\tilde{R}(\lambda) = \|\Pi \hat{a}_\lambda\|_2^2 + 2(\text{trace}(\Pi M_\lambda) - \hat{a}_\lambda^\top M^\dagger b). \quad (15)$$

While introducing bias into the estimator of statistical risk seems bad, it can be necessary in situations where M is ill-conditioned. Small eigenvalues of M corresponds to there being very little information along the associated eigenvectors. By zeroing out these eigenvalues we are effectively saying we cannot reliably estimate with this little amount of information. Hence, we concentrate on getting the estimation correct on the eigenvectors with larger eigenvalues.

For each object detected in the CFHT images, and for given values of condition number Λ ranging from $4 \leq \log(\Lambda) < 6$, we evaluate $R(\lambda)$ at values of $-2 \leq \log(\lambda) < 2$. Figure 4 shows a typical outcome of this analysis for

⁶ It should be noted that other risk estimators may be constructed, e.g. ones that maximize the quality of the full difference image.

a single object. Along the y-axis we show the associated value of the conditioning parameter Λ , and along the x-axis the value of λ at which $R(\lambda)$ is evaluated. The *solid* line shows the minimum value of $R(\lambda)$ for each Λ .

We note that as we decrease the acceptable matrix condition number, thereby truncating more eigenvalues from the matrix pseudoinverse, the optimum value of λ increases. For matrices with effectively no conditioning (large Λ), the optimal value of λ is near $\lambda = 0.5$. This is in fact the global minimum of the risk. A similar result is obtained by looking at all objects within an image and summing their cumulative risk surfaces. We regard $\lambda = 0.5$ as the value preferred by the empirical risk estimation technique, with a range of nearly-equivalent risk between $0.3 < \lambda < 1.0$.

5.1.2. Predictive Ability

In most PSF-matching implementations, several dozen objects across a pair of registered images are used to create individual $K(u, v)$; ideally these should evenly sample the spatial extent of the images. Due to spatial variation in the PSFs of the images, caused by optical aberrations or bulk atmospheric effects, the single kernel that PSF-matches *all* objects in an image must itself vary spatially. In this case each of the kernels $K(u, v)$ are used to build spatially varying PSF-matching kernel $K(u, v, x, y)$. This is typically implemented as spatial variation on the kernel coefficients $K(u, v, x, y) = \sum_i a_i(x, y) K_i(u, v)$. Therefore an additional consideration in the choice of λ is the ability to build spatial models for the coefficients $a(x, y)$.

To quantify this, we examine the predictive ability of the kernel solution \hat{a}_λ . In all CFHT images, we identify object pairs separated by more than 5 pixels but less than 50, a range of separations where we expect the intrinsic spatial variation of the underlying kernel to be minimal. The kernel derived for each object in a pair is applied to its complement, and the quality of each difference image assessed. For components A and B of each object pair, this yields difference image D_{AA} which is the difference image of object A with kernel A, D_{AB} which is the difference image of object A with kernel B, and analogous images D_{BA} and D_{BB} . We assess the quality of each difference image using the width of the pixel distribution normalized by the noise, defined as e.g. σ_{AA} , within the central 7×7 pixels of the difference image. While we don't expect this distribution to have a width of exactly 1.0 due to covariance between the solution and the input images, we do desire that the quality of D_{AB} and D_{BA} should not be significantly worse than that of D_{AA} and D_{BB} .

We aggregate the “even” statistics σ_{AA} and σ_{BB} into distribution Σ_E , and the “odd” statistics (σ_{AB} , σ_{BA}) into Σ_O . We further examine the distribution of Σ_{O-E}^2 , which is created from all measurements of $\sigma_{AB}^2 - \sigma_{AA}^2$ and $\sigma_{BA}^2 - \sigma_{BB}^2$. This statistic reflects the deterioration in an object's difference image when using a counterpart's kernel, compared to the optimal kernel derived for that object.

We plot the distributions of these values in Figure 5. The *top* panel provides the median values of these distributions for the sum-of-Gaussian (AL) basis (*left*), for the unregularized delta function basis ($\lambda = 0$; *cen-*

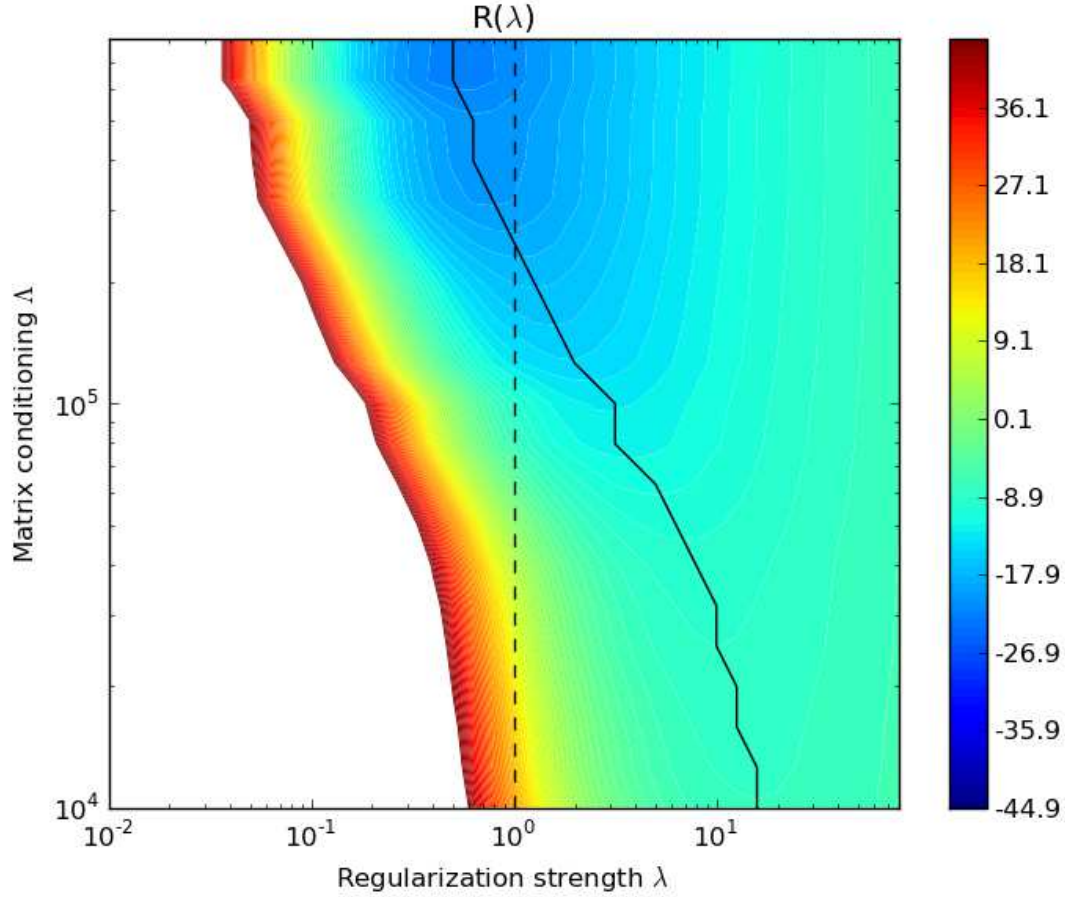


Figure 4. Values of the empirical risk $R(\lambda)$, as defined in Equation 14, for different values of the matrix conditioning parameter Δ , and the regularization strength λ . At all Δ , we determine the minimum values of $R(\lambda)$, which are connected by the *solid* black line. The *dotted* vertical line represents the fiducial value of $\lambda = 1$. The global minimum of $R(\lambda)$ is realized with minimal matrix conditioning, and at a value of $\lambda = 0.5$.

ter), and for delta-function regularization strengths of $-2 < \log(\lambda) < 2$ (right). The *bottom* panel plots the effective standard deviation of the distribution, defined as 74% of the interquartile range.

The lowest median residual variance Σ_E comes from difference images made using an unregularized $\lambda = 0$ basis, the reasons for which we have examined in detail in Section 4. However, as expected the predictive ability of this basis is by far the worst, having the highest median Σ_{O-E}^2 , as well as large variance within this distribution. As we ramp up the regularization strength, the predictive ability of the kernels increases (low Σ_{O-E}^2), but at the expense of the quality of the difference image itself (large Σ_E).

To find an acceptable medium between these two considerations, we will use the results from the sum-of-Gaussian (AL) basis as a benchmark, since it has been shown to produce effective spatial models (Section 3). For the AL basis, the median values of Σ_E , Σ_O , and Σ_{O-E}^2 are 0.99, 1.14, and 0.28, respectively. Similar results are obtained with delta-function regularization strengths of $\lambda \approx 0.2, 0.7$, and 0.2. For AL the σ_{median} values of Σ_E , Σ_O , and Σ_{O-E}^2 are 0.14, 0.33, and 0.74, respectively. These are matched (or bested) in the reg-

ularized basis for $\lambda \leq 0.2$, $\lambda = 0.2$, and $0.2 \leq \lambda \leq 6$, respectively.

In summary, using delta-function regularization strengths of $\lambda \approx 0.2$, we are able to achieve difference images with a similar quality to those yielded by the sum-of-Gaussian AL basis (using Σ_E as our metric). These models have similar predictive ability when applied to neighboring objects (quantified using Σ_O and Σ_{O-E}^2), making them useful for full-image spatial modeling. Finally, they are seen to be generally applicable, having a small variance in the above statistics when evaluated over several hundreds of object pairs.

6. CONCLUSIONS

We’ve examined here the choice of basis set on the quality of PSF-matching kernels and their resulting difference images. These include the traditional sum-of-Gaussian (“Alard-Lupton”) basis and a digital basis based upon delta-functions. We find that while the delta-function kernels are the most expressive, they are also the least compact in terms of localization of power within the kernel. Having one basis component per pixel in the kernel, they tend to overfit the data and are more sensitive to the noise in the images instead of the intrinsic

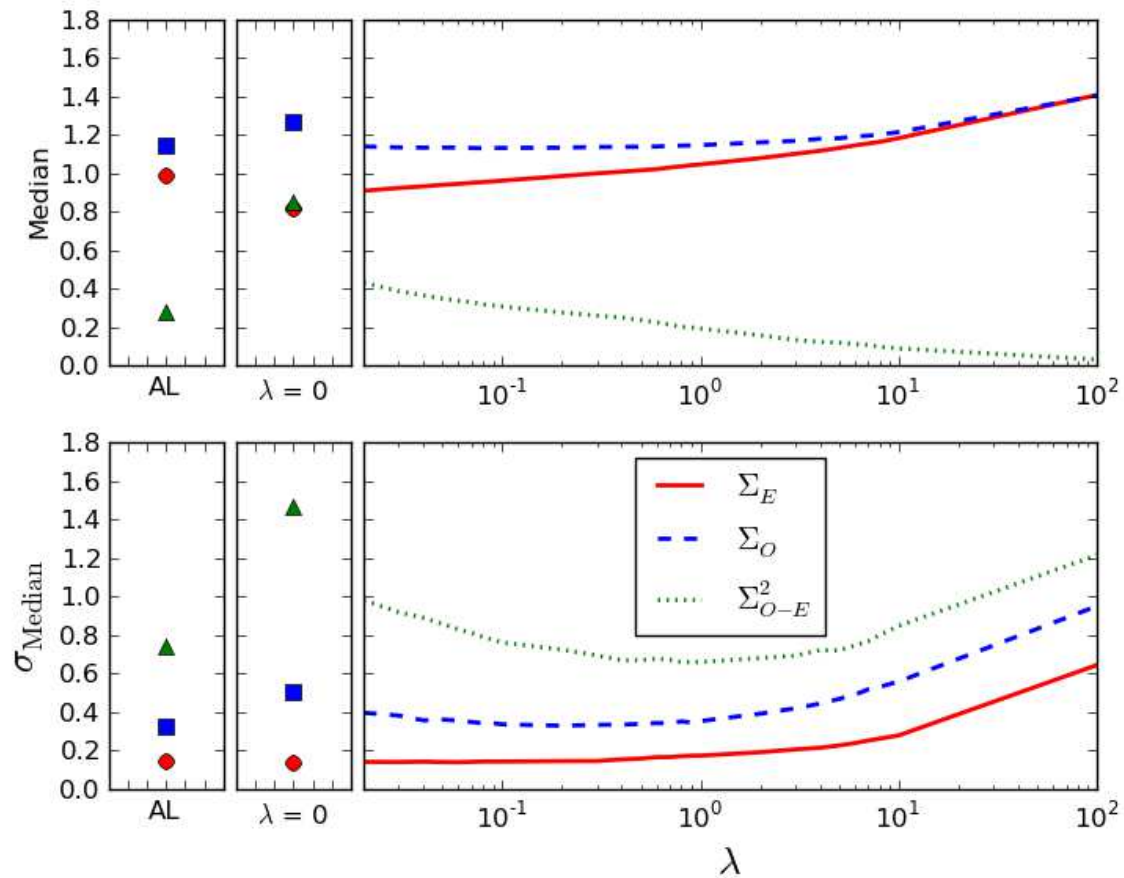


Figure 5. Median statistics assessing the predictive ability of different kernel bases. The top panel shows the median values of statistics Σ_E (red circle and solid line), Σ_O (blue square and dashed line), and Σ_{O-E}^2 (green triangle and dotted line) for “Alard-Lupton” (AL) bases, for delta-function bases with $\lambda = 0$, and then for a range of $-2 < \log(\lambda) < 2$. All statistics are defined in Section 5.1.2. The bottom panel shows the standard deviation of the distribution, defined as 74% of the interquartile range.

PSF-matching signal.

We introduce a new technique of linear regularization to impose smoothness on these delta-function kernels, at the expense of slightly higher noise in the difference images. These regularized shapes are shown to be flexible, and yield solutions with sufficient predictive power to prove useful for spatial interpolation. We outline two methods to determine the strength of this regularization that minimize the statistical risk of the kernel estimate, and that examine the predictive ability of the derived kernels. Both methods suggest values of λ that are between 0.1 and 1.0.

Given the large range of image qualities used in image subtraction pipelines compared to the small number of images used in the analysis here, we caution that these estimates may not be applicable under all conditions and should really be estimated on a dataset-by-dataset basis. The optimal value of λ will be a function of the S/N in the template and science images, which should affect the level of kernel smoothing needed, and of the respective seeings in the input images, which may impact the suitability of our finite-difference smoothness approximation.

While this implementation appears successful and

practical, there are various improvements we might consider in our regularization efforts. This includes changing the scale over which the regularization stencil is calculated based upon the seeing in the images; currently this is being done in pixel-based coordinates, and not adjusted depending on the full-width at half-maximum of the input PSFs. We also plan to examine additional metrics to determine the optimal value of λ , including the power spectrum of noise in the resulting difference image, which should be flat. Ultimately, the overall quality of the *entire* difference image is the optimal metric to use in assessing choice of basis; we will be expanding our analysis to include full-image metrics and spatial modeling of the kernel.

Finally, the wealth of statistical techniques to efficiently choose basis shapes has not been exhausted. Other potential methods include the use of overcomplete bases, where the choice of the correct subset of components to use is made through basis pursuit (Chen et al. 1998), as well as the process of “basis shrinkage” through the use of multi-scale wavelets (Donoho & Johnstone 1994, 1995). In all considerations, it is an advantage to yield solutions that, as an ensemble, have a low dimensionality so that spatial modeling is ef-

ficient and spatial degrees of freedom are not being used to compensate for an inefficient choice of basis. However, for any given basis set the choice of regularization (none at all or using a fixed set of functions) is likely to be the proper place for optimization.

This material is based, in part, upon work supported by the National Science Foundation under Grant Number AST-0709394.

REFERENCES

- Alard, C. 2000, *A&AS*, 144, 363
 Alard, C. & Lupton, R. H. 1998, *ApJ*, 503, 325
 Albrow, M. D., et al. 2009, *MNRAS*, 397, 2099
 Alcock, C., et al. 1999, *ApJ*, 521, 602
 Becker, A. C., et al. 2004, *ApJ*, 611, 418
 Bond, I. A., et al. 2001, *MNRAS*, 327, 868
 Botticella, M. T., et al. 2010, *ApJ*, 717, L52
 Bramich, D. M. 2008, *MNRAS*, 386, L77
 Chen, S. S., Donoho, D. L., Michael, & Saunders, A. 1998, *SIAM Journal on Scientific Computing*, 20, 33
 Darnley, M. J., et al. 2007, *ApJ*, 661, L45
 Donoho, D. L. & Johnstone, I. M. 1994, *Biometrika*, 81, 425
 —. 1995, *Journal of the American Statistical Association*, 1200
 Israel, H., Hessman, F. V., & Schuh, S. 2007, *Astronomische Nachrichten*, 328, 16
 Kerins, E., et al. 2010, *MNRAS*, 409, 247
 Kohavi, R. 1995, in (Morgan Kaufmann), 1137–1143
 Magnier, E. A. & Cuillandre, J. 2004, *PASP*, 116, 449
 Miknaitis, G., et al. 2007, *ApJ*, 666, 674
 Miller, J. P., Pennypacker, C. R., & White, G. L. 2008, *PASP*, 120, 449
 Newman, A. B. & Rest, A. 2006, *PASP*, 118, 1484
 Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. 1992, *Numerical recipes in C. The art of scientific computing*
 Rest, A., et al. 2005, *ApJ*, 634, 1103
 Sako, M., et al. 2008, *AJ*, 135, 348
 Smith, C., et al. 2002, in Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference, Vol. 4836, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, ed. J. A. Tyson & S. Wolff, 395–405
 Stein, C. M. 1981, *The Annals of Statistics*, 9, 1135
 Tomaney, A. B. & Crotts, A. P. S. 1996, *AJ*, 112, 2872+
 Udalski, A., Szymanski, M. K., Soszynski, I., & Poleski, R. 2008, *Acta Astronomica*, 58, 69
 Wozniak, P. 2008, in Manchester Microlensing Conference
 Wozniak, P. R. 2000, *Acta Astronomica*, 50, 421
 Wnsche, A. 2000, *Journal of Physics A: Mathematical and General*, 33, 1603